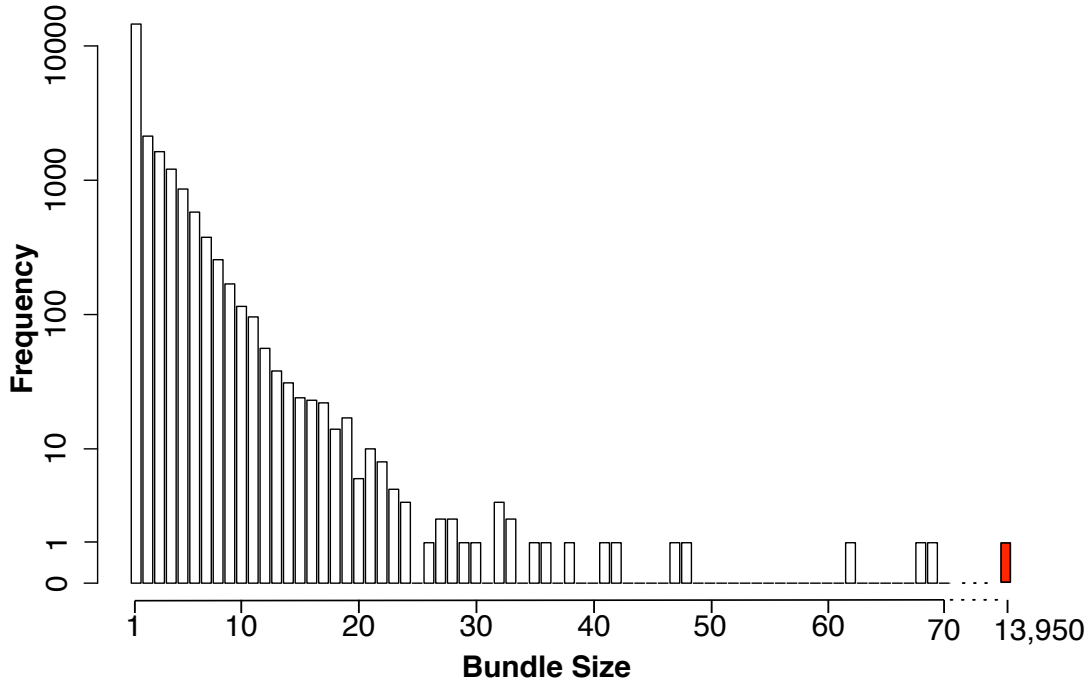
STREAMING FRAGMENT ASSIGNMENT FOR REAL-TIME ANALYSIS OF SEQUENCING EXPERIMENTS

ADAM ROBERTS AND LIOR PACHTER

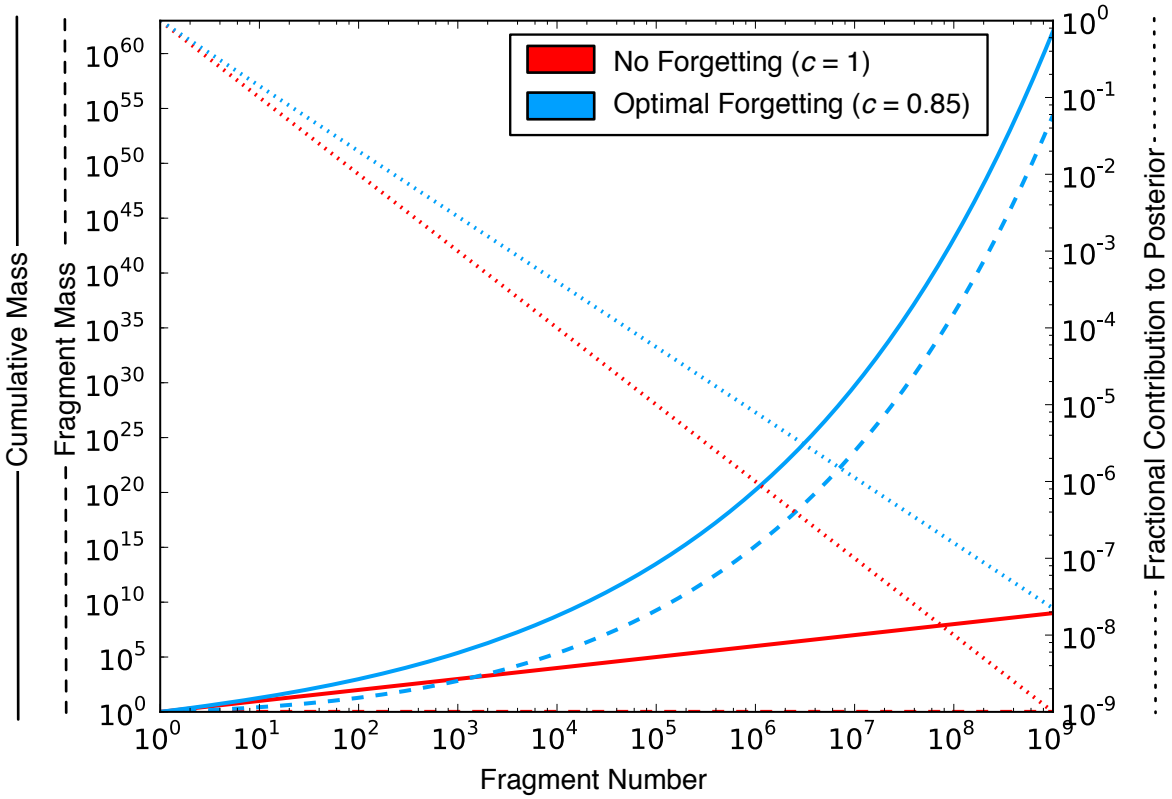
Supplementary Figure 1



Supplementary Figure 1: Histogram of bundle sizes. The bundle graph has target sequences as vertices and an edge between two vertices if there is a fragment that maps ambiguously to both. A bundle is a connected component of the bundle graph. The histogram shows bundle sizes for the simulation with 1 billion reads, at which point the largest bundle has 13,950 transcripts (highlighted in red). With the batch algorithm, all reads mapping to these transcripts must be processed simultaneously, leading to a major computational bottleneck. This effect can be viewed as an instance of the "curse of deep sequencing" which describes the phenomenon that as more reads are sequenced, bundle sizes grow due to errors in reads that lead to spurious mappings. `eXpress` avoids this problem via the online algorithm.

Nature Methods: doi:10.1038/nmeth.2251

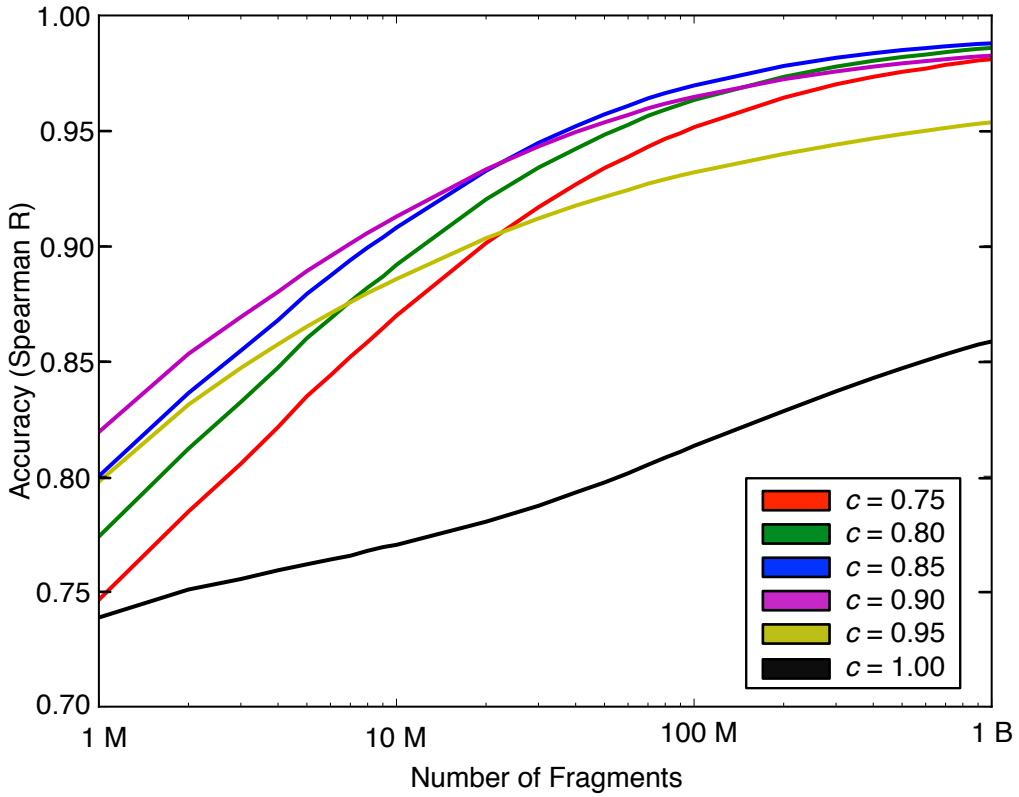
Supplementary Figure 2



Supplementary Figure 2: Growth of the forgetting factor. In the simplest implementation of the online algorithm, all incoming fragments are given a mass of 1, corresponding to a forgetting factor of 1. However, as the cumulative mass grows, the fragment mass does not, and later fragment assignments have progressively smaller influence on the posterior distribution. By increasing the mass of later fragments using a forgetting factor, the fragment mass grows with the cumulative mass to reduce the effect of the prior. This allows for much faster convergence in practice, but can also cause instability if the mass grows too quickly. In this plot, we show the increase in fragment mass for a forgetting factor of 0.85, which we found empirically to give the best results in our simulations (see Supplementary Figure 3).

Nature Methods: doi:10.1038/nmeth.2251

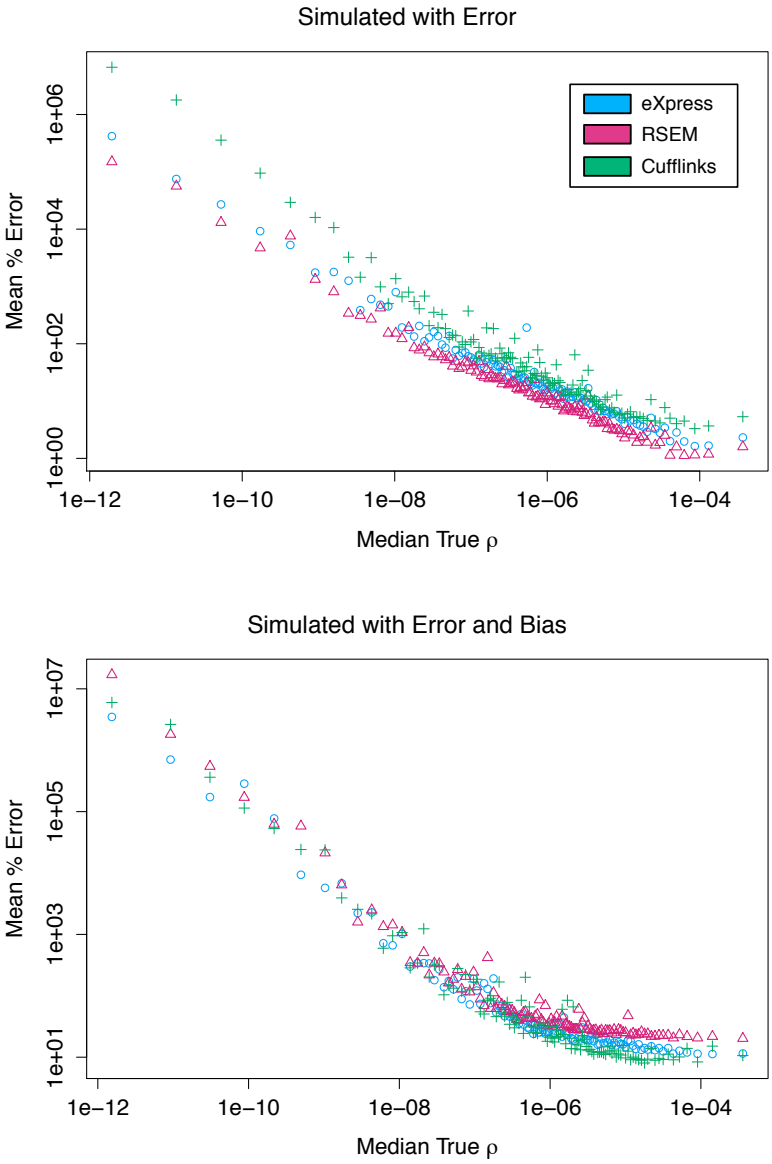
Supplementary Figure 3



Supplementary Figure 3: Impact of forgetting factor on accuracy and stability. This plot shows the accuracy of `eXpress` for different forgetting factors (values of c) on the unbiased simulation data (see Methods). The solid blue line for $c = 0.85$ was selected as optimal and corresponds to the solid blue line in Figure 2a. A forgetting factor value of $c = 1.00$ corresponds to the scenario where all fragments are given an equal weight (see Supplementary Figure 2).

Nature Methods: doi:10.1038/nmeth.2251

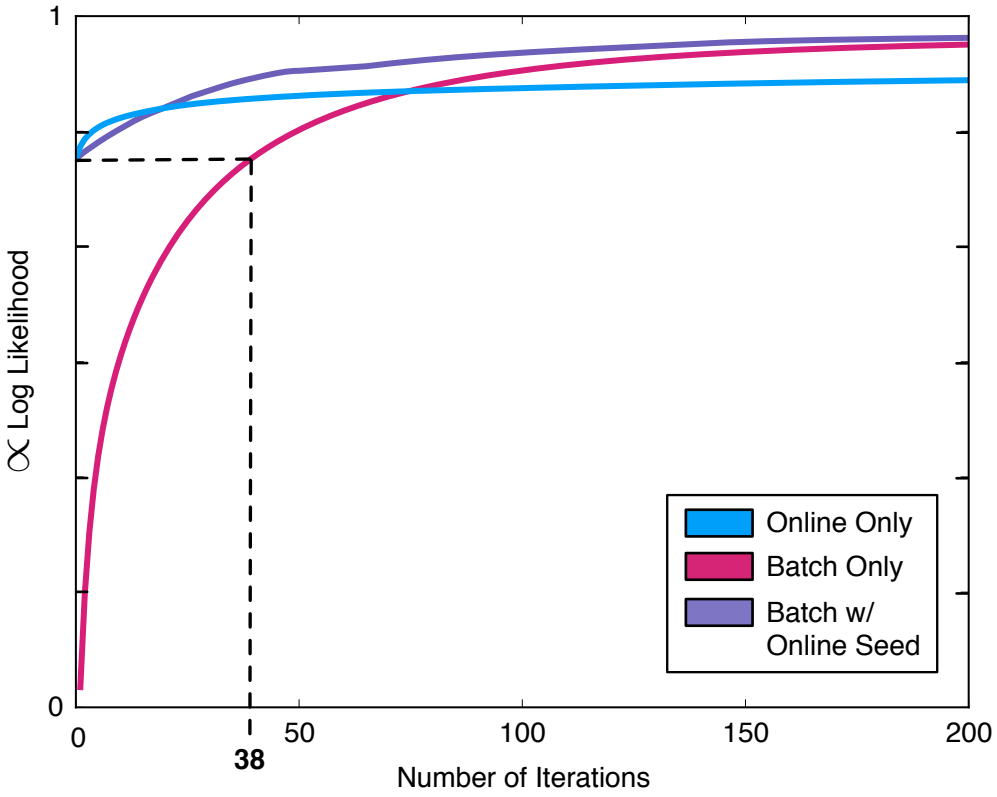
Supplementary Figure 4



Supplementary Figure 4: Estimation accuracy at different abundances. Using the simulated reads from Figure 2 with (bottom) and without (top) bias, the transcripts were separated into bins of 500 targets by the true (simulated) abundance level (ρ). For each bin, the mean percent error of the estimated ρ is shown for all three tools. When no bias is included, the results mirror those observed in Figure 1, which shows Spearman correlation as a summary for all abundance levels. However, in contrast to the Spearman correlation results in Figure 2, Cufflinks has lower error than eXpress for a majority of the bins when bias is included. Both figures use estimates from 500 million reads. Note that both axes are on the log scale.

Nature Methods: doi:10.1038/nmeth.2251

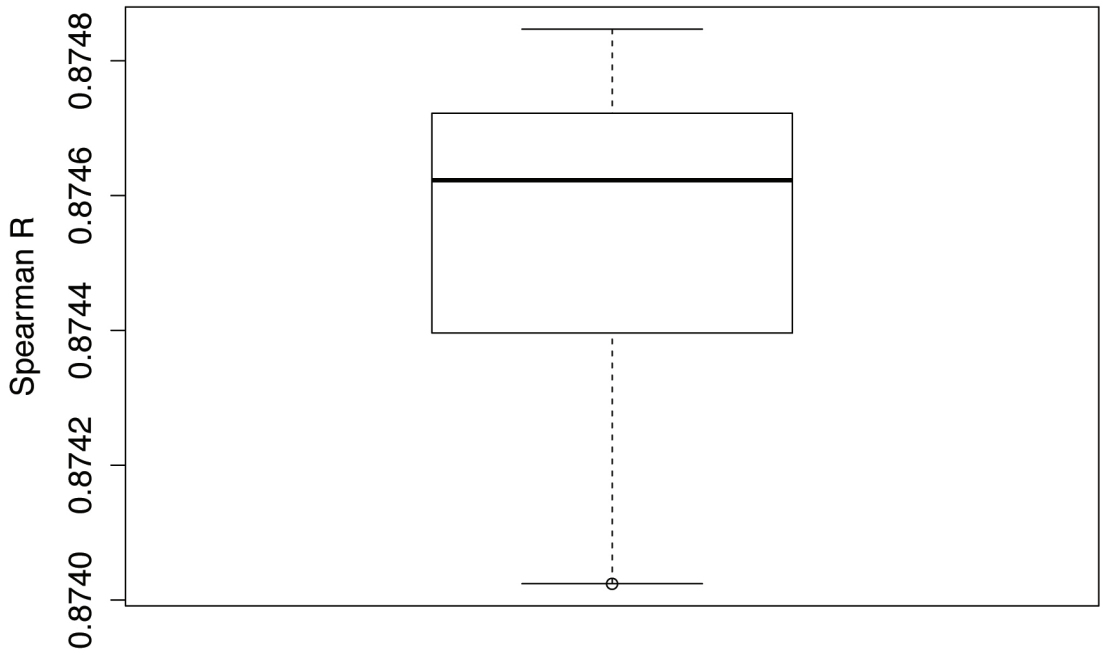
Supplementary Figure 5



Supplementary Figure 5: Comparison of different iterative methods. The log likelihood was calculated using our model (see Methods) on a set of 25 million simulated pair-end reads. The batch only curve (red) shows the log likelihood achieved using the standard batch algorithm where in each round all fragments are incorporated in the expectation step of the EM algorithm. The online curve (blue) shows the log likelihood achieved by repeated passes through the data as if it were additional observations (i.e., the weighting is not reset for each round). The purple curve shows a coupled method where the online algorithm is used to "seed" the parameters for the batch algorithm. The 0^{th} iteration corresponds to this seed round. Note that in this experiment, a single round of the online algorithm yields results equivalent to 38 rounds of the batch algorithm. The "batch with online seed" has superior performance to the other methods after 21 rounds.

Nature Methods: doi:10.1038/nmeth.2251

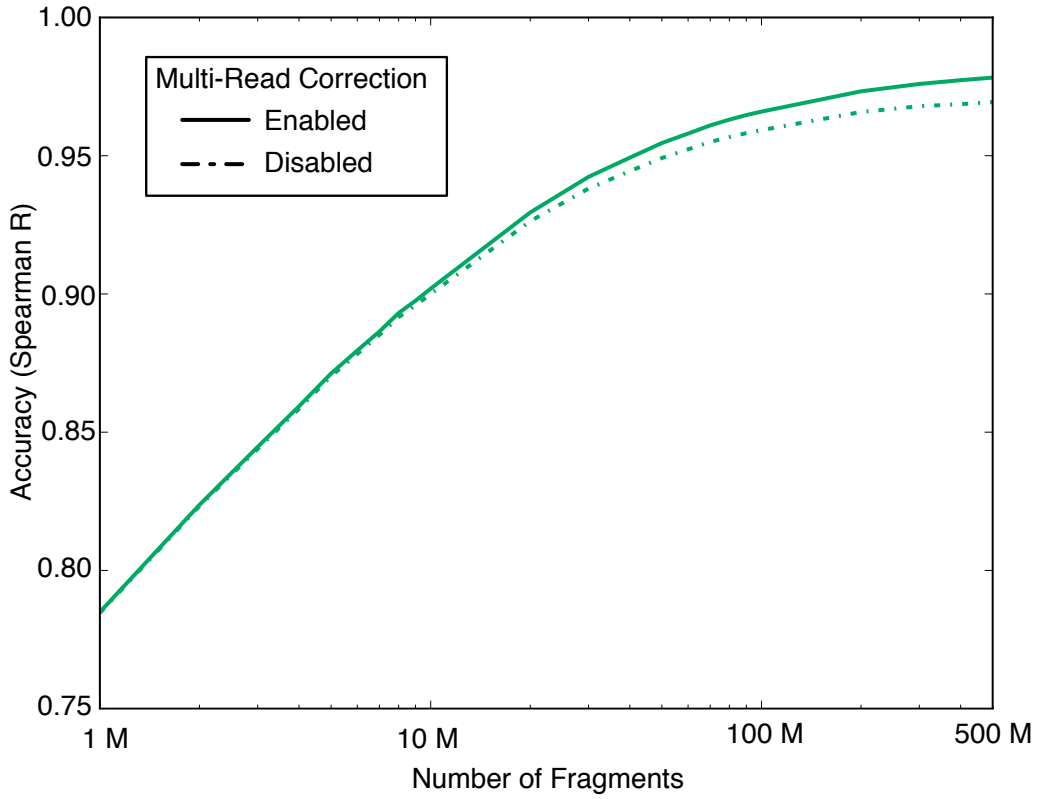
Supplementary Figure 6



Supplementary Figure 6: Effect of read order. Convergence of the online EM algorithm is guaranteed only when the data is randomly ordered. To determine whether RNA-Seq reads are randomly ordered as needed, the 33.2 million paired-end reads from the ENCODE dataset used in the main text was randomly shuffled 9 times. For each random shuffle as well as the original ordering, the mean Spearman correlation of abundance estimates with all other orderings was computed, and the distribution is shown above. With a variance of 6.3×10^{-8} we conclude that, when randomized, the ordering has practically no effect on abundance estimation. Furthermore, the unshuffled order output by the Illumina sequencer (represented by the circle) is within 2 standard deviations and can be assumed random.

Nature Methods: doi:10.1038/nmeth.2251

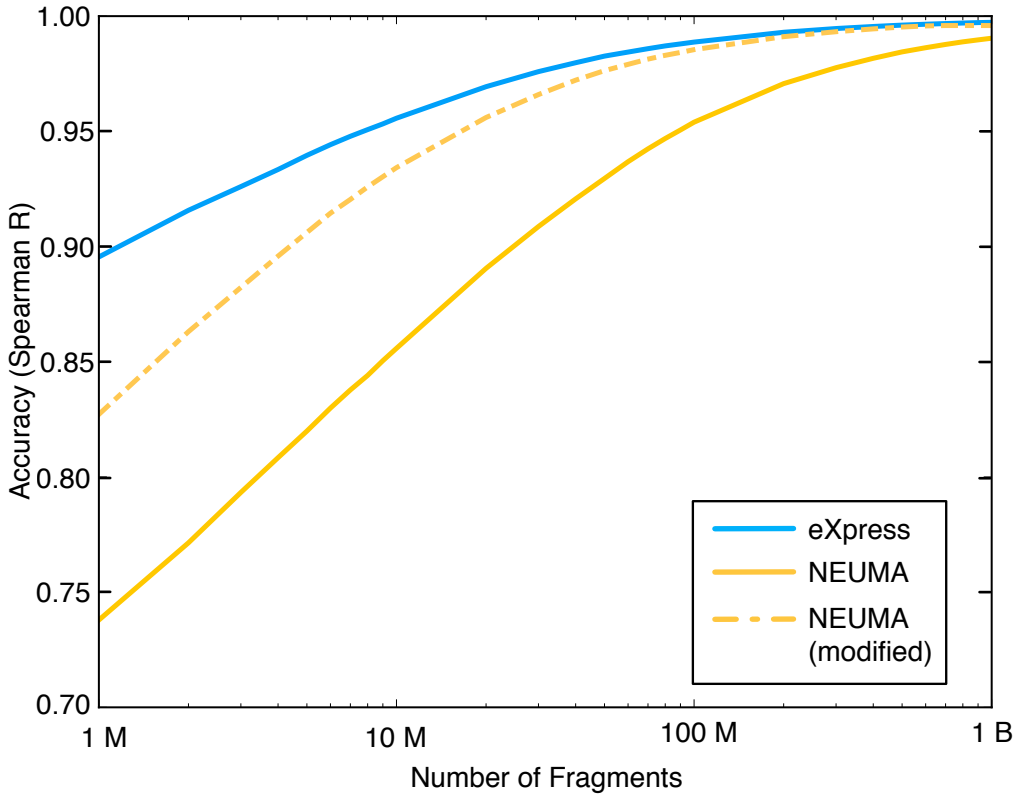
Supplementary Figure 7



Supplementary Figure 7: Effect of multi-mapping reads on `Cufflinks` accuracy. `Cufflinks` reduces its time and memory requirements by assuming each set of overlapping transcripts to be independent of all others. Thus, a heuristic is used to disambiguate fragments that map to multiple genomic locations. The original `Cufflinks` heuristic (before v1.0, dashed line) distributed multi-mapping reads equally amongst the loci they mapped to. We modified the software to instead distribute the reads according to the `Cufflinks` model following initial abundance estimation that uses a uniform distribution, which is equivalent to a single round of EM over the multi-mapping reads. The results show improvement at sufficiently high sequencing depth (solid line). This approach was used for all results in our paper and can be enabled with the -u option in `Cufflinks`.

Nature Methods: doi:10.1038/nmeth.2251

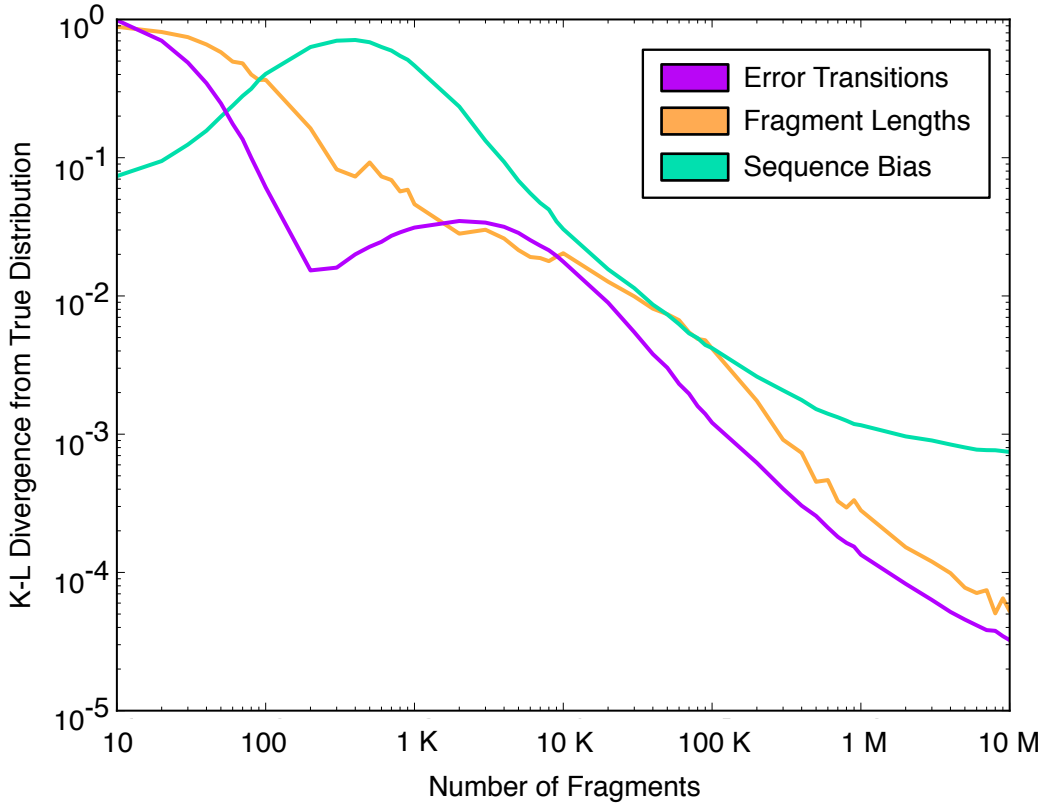
Supplementary Figure 8



Supplementary Figure 8: Limits of quantification when ignoring ambiguous reads. `NEUMA` (Normalization by Expected Uniquely Mappable Areas) [11] calculates an effective length for each transcript in order to normalize counts based on uniquely mappable areas of transcripts. We modified `NEUMA` to allow for errors (see Methods), thereby increasing the accuracy of the method considerably, but its accuracy remains inferior to `eXpress`, which does consider ambiguous reads. Furthermore, `NEUMA` is unable to produce abundance estimates for targets without sufficient amounts of unique sequence. The EM algorithm is superior because it can take advantage of different combinations of shared sequence among multiple targets to produce estimates. The accuracy was calculated as described in the Methods, using only the subset of targets (77% of total) that `NEUMA` quantifies.

Nature Methods: doi:10.1038/nmeth.2251

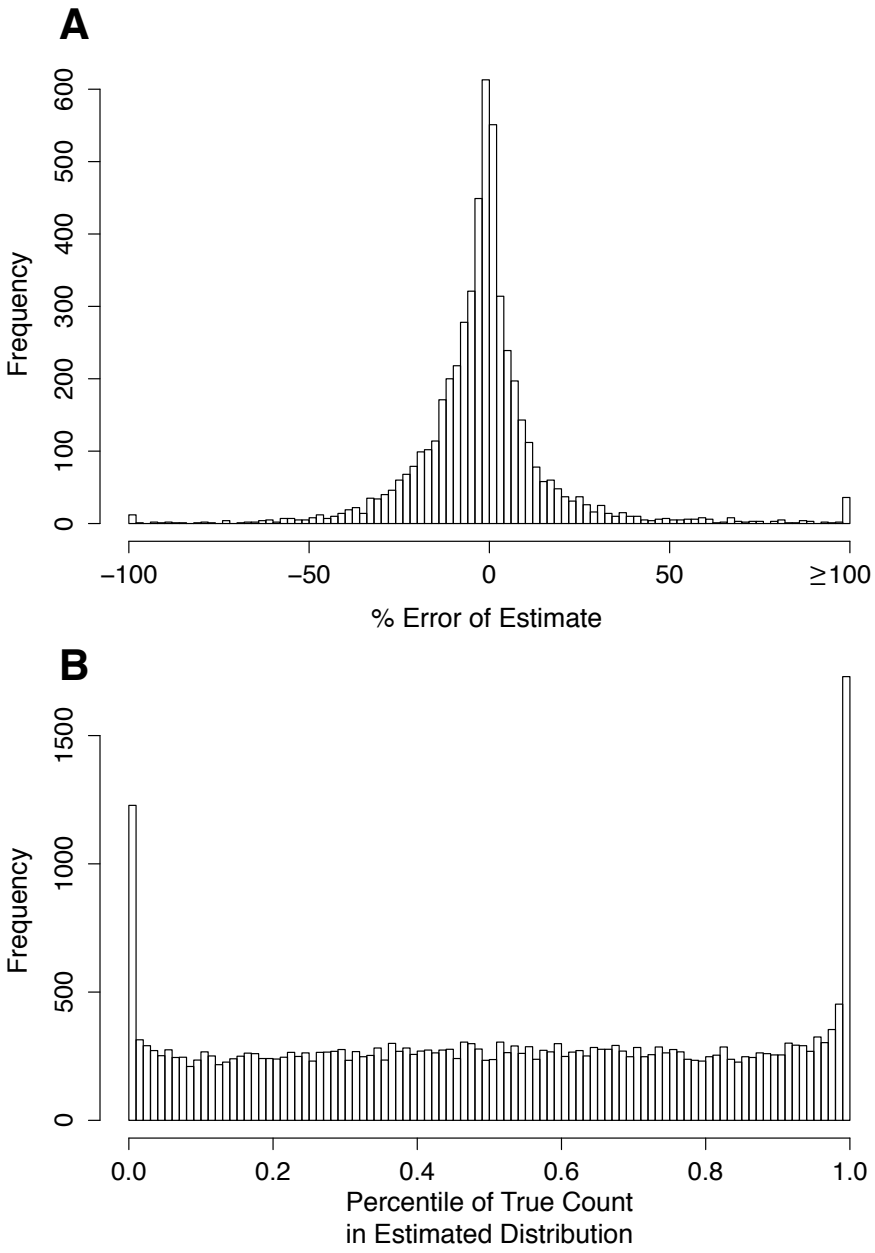
Supplementary Figure 9



Supplementary Figure 9: Convergence of auxiliary parameters. Convergence of auxiliary parameter estimates was measured by Kullback-Leibler divergence from the true distributions. In our simulations all have a divergence below 10^{-3} by 2 million observed fragments, which equates to a an average (weighted) relative ratio between the estimate and the truth of 1.001. Therefore, in our implementation we fix the auxiliary parameters and stop learning after 5 million fragments. Sequence bias converges more slowly due to other approximations used in our implementation (see Methods).

Nature Methods: doi:10.1038/nmeth.2251

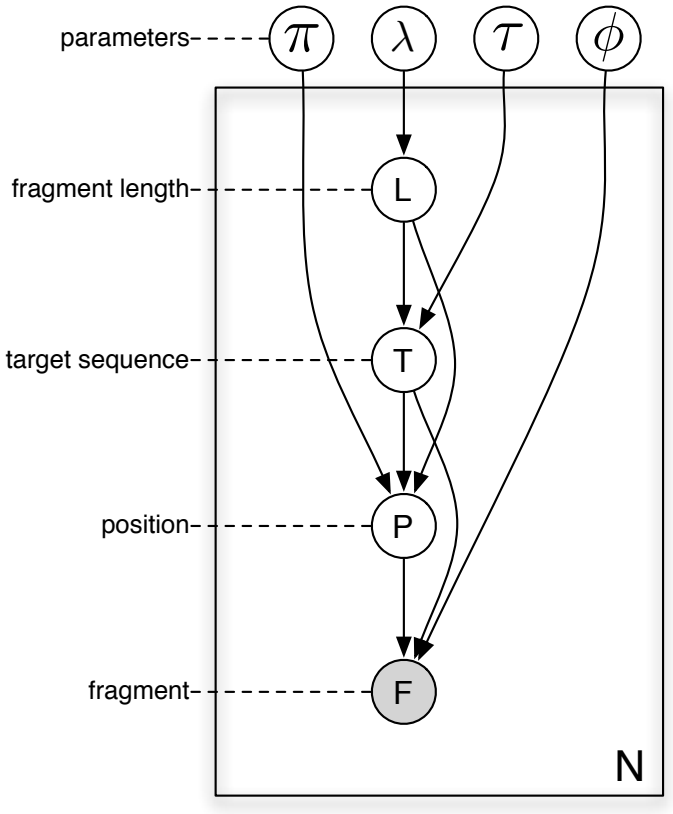
Supplementary Figure 10



Supplementary Figure 10: Accuracy of count posterior distributions. Counts were estimated using the first 25 million fragments in the dataset simulated without bias. (a) Histogram of percent error for estimated ambiguous counts on targets marked as solvable with an FPKM of at least 1. (b) Histogram showing frequencies of the percentile of the true count in the estimated shifted beta-binomial distribution for each target marked solvable.

Nature Methods: doi:10.1038/nmeth.2251

Supplementary Figure 11



Supplementary Figure 11: The `eXpress` model. A graphical model describing the generative process for obtaining fragments that underlies the online EM algorithm used by `eXpress`. The model represents the relationship between the sequence bias (π), fragment length distribution (λ), target abundances (τ), and error transition probabilities (ϕ) that produce the observed fragment sequences.

Nature Methods: doi:10.1038/nmeth.2251

Supplementary Table 1

Method	w/o Bias Correction	w/ Bias Correction
`eXpress`	0.807	0.834
`RSEM`	0.791	-
`Cufflinks`	0.797	0.836

Supplementary Table 1: Validation with qRT-PCR. Spearman's rank correlation coefficients for comparisons between tested methods' abundance estimates and qRT-PCR for 907 transcripts measured by MAQC, as described in [10]. While all methods have approximately the same accuracy, `eXpress` and `Cufflinks` benefit from the bias correction method described here and in [10]. These results are concordant with the improvements due to bias correction reported in [4].

Nature Methods: doi:10.1038/nmeth.2251

Supplementary Table 2

Method	# Mismatches Allowed	# Unmapped Pairs	# Total Alignments	eXpress Runtime (min)	ρ_{est} Accuracy (Spearman R)
Hobbes	0	33548108	62921224	31.23	0.880
Hobbes	1	4450871	223530695	80.91	0.921
Hobbes	2	309844	324849991	111.61	0.931
Hobbes	3	23952	360692805	121.17	0.932
Hobbes	4	11766	272735164	128.75	0.932
Hobbes	5	11408	379299189	127.55	0.932
Hobbes	6	11393	385802725	132.99	0.931
Hobbes	7	11391	394820807	132.15	0.930
Hobbes	8	11386	410043522	142.30	0.929
Hobbes	9	11386	437999906	150.53	0.928
Bowtie	3	10646188	323058366	121.91	0.934

Supplementary Table 2: Performance with different mappers and settings. 100 million of the unbiased simulated read set were mapped with Hobbes [15] using different numbers of allowed mismatches per read (not pair) and compared with Bowtie [14] mapping allowing for 3 mismatches (the max). The Bowtie mappings produce fewer alignments at 3 allowed mismatches than Hobbes because these mismatches are constrained to occur beyond the "seed", which must match perfectly. When no mismatches are allowed, less than 36% of the pairs align to the transcriptome, each with approximately 2 mappings. This allows eXpress to process the data in approximately 30 minutes, but produces a relatively low Spearman correlation ($R = 0.88$). The accuracy peaks at $R = 0.9324$ when 4 mismatches are allowed, and then slowly decreases as improper alignments begin to be more prevalent. For example, allowing 9 mismatches leads to 4.38 mappings per pair, most of which are presumably due to random chance. Thus, we find it is important to allow a sufficient number of mismatches, while it is reassuring that the speed and accuracy of eXpress do not degrade significantly when many more mismatches are allowed.

Nature Methods: doi:10.1038/nmeth.2251